Progression of Parkinson's Disease using UPDRS and machine learning algorithms on proteins and peptides features

Major project dissertation submitted to the University of Delhi in partial fulfillment of the requirements for the award of the degree of

MASTER OF SCIENCE (COMPUTER SCIENCE)

by

VISHVENDRA SINGH

21234747061

Under the supervision of

Dr. Bharti
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF DELHI
DELHI-110007, INDIA

July 2023

Certificate

This is to certify that the major project dissertation entitled **"Progression of Parkinson's Disease using UPDRS and machine learning algorithms on proteins and peptides features"** being submitted by **Vishvendra Singh** under the supervision of **Dr. Bharti,** Assistant Professor, Department of Computer Science, University of Delhi. The project work has been carried out for the partial fulfillment of the requirements of Master of Science (Computer Science) degree in the Department of Computer Science, University of Delhi. The project embodies original research work carried out at the Department of Computer Science, University of Delhi. This has not been submitted so far, in part or full, to any other University or institute for the award of any other degree or diploma.

Vishvendra Singh 21234747061

Dr. Bharti
Supervisor
Department of Computer Science
University of Delhi
Delhi-110007

Prof. Naveen Kumar
Head
Department of Computer Science
University of Delhi
Delhi-110007

Declaration

We hereby declare that the major work project dissertation entitled **"Progression of Parkinson's Disease using UPDRS and machine learning algorithms on proteins and peptides features"** which is being submitted to Department of Computer Science, University of Delhi, Delhi-110007 in the partial fulfillment of the requirements of Master of Science (Computer Science) degree is a bonafide work carried out by us. The work has been carried out under the supervision of Dr. Bharti, Assistant Professor, Department of Computer Science, University of Delhi.

The project embodies original work carried out at the Department of Computer Science. This work has not been submitted, in part or full, to any other University or Institute for the award of any other degree or diploma.

Vishvendra Singh
21234747061

Acknowledgment

We would like to extend our profound gratitude to our supervisor, Dr. Bharti for her interest, guidance, and valuable suggestions throughout the course of this project. We feel honored and privileged to work under her constant supervision. Without her advice, constructive ideas, positive attitude, and continuous encouragement, it would not have been possible to make such progress in the designated time frame.

"Data used in the preparation of this article were obtained from a Kaggle competition(www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction/data).

Kaggle is an online platform and community that brings together data scientists, machine learning engineers, and data enthusiasts from around the world. It was founded in 2010 and has since become one of the largest and most popular platforms for data science competitions and collaboration. One of the key features of Kaggle is its data science competitions. Organizations and companies post real-world datasets and problem statements, inviting participants to develop innovative solutions and models. In addition to competitions, Kaggle hosts a vast collection of datasets that users can explore and analyze.

Vishvendra Singh
21234747061

Abstract

Parkinson's disease (PD) is a neurological illness characterised by discomfort, tremor, and slowness of movement. The quality of life in terms of health has been proven to be considerably impacted by Parkinson's disease.

Its early and accurate diagnosis is a prime concern for clinicians and patients. This project aims to develop an automated system to predict updrs_score with machine learning approaches. Further, if any disease impacts protein and peptides, we aim to learn a regression model using uniport and peptide as features.

In the present work, We constructed three datasets containing different numbers of features from the datasets given on Kaggle online competition “AMP-Parkinson's Disease Progression prediction”. Support Vector Regression, Random Forest Regression, and MultiLayer Perceptron were three machine learning models that we worked with.

Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R_Score (r2) are used to evaluate the performance utilising the k-fold cross-validation method. The best r_score of 0.51±0.00 was achieved using Random Forest. We were also able to identify the important regions. This study may be extended further in future.

Table of contents

1. Introduction

1.1. Background

Parkinson's disease is a brain condition that results in uncontrolled or unintentional tremors, slowness, and balance and coordination issues. Most frequently, symptoms develop gradually and worsen with time. Moving around and communicating may become challenging as the sickness becomes worse. They might also go through mental and behavioral changes, memory loss, sleep problems, depression, and weariness. Parkinson's disease has been defined by the loss of nerve cells in the movement-regulating region of the brain, the basal ganglia, which results in the disease's most evident symptoms. Typically, these nerve cells, or neurons, create dopamine, a crucial neurotransmitter in the brain. The condition causes problems with movement because the degeneration or death of the neurons reduces the amount of dopamine that is produced. What leads to the degeneration of neurons is yet unknown to scientists.(“Parkinson’s Disease: Causes, Symptoms, and Treatments”).

Symptoms include

- Slow movement
- Trembling in the jaw, head, arms, or legs
- Imbalance and coordination issues, which can occasionally cause falls
- Muscular stiffness caused by persistently tightened muscles

1% of people over the age of 65 years suffer from PD (estimated about 12 lac Indians living with PD in India) (“Parkinsons Disease and the Ageing Indian Population - Healthcare Radius” 2021). Parkinson's disease is predicted to affect 1.6 million Americans by 2037, costing the country close to $80 billion.(“AMP®-Parkinson’s Disease Progression Prediction | Kaggle”).

1.2. Problem statement

“The Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale” (MDS-UPDRS) provides a comprehensive assessment of both motor as well as non-motor symptoms that correspond to Parkinson's.

There are four types of updrs scores(Goetz et al. 2008) -

Updrs_1 : This section of the scale evaluates how Parkinson's disease affects patients' everyday experiences in non-motor domains like depression, anxiety, apathy, difficulty sleeping, pain and other physical symptoms, and exhaustion.

Updrs_2 : This section of the scale evaluates how Parkinson's disease affects patients' everyday experiences in motor domains like speech, freezing, walking and balance, tremor, handwriting, dressing, and chewing and swallowing.

Updrs_3 : This section of the scale evaluates the motor symptoms of Parkinson's disease like speech, postural stability, leg agility, toe tapping, hand movements, finger tapping, and facial expression.

Updrs_4 : In this section, the rater assesses the motor issues based on historical and objective data, such as the span of dyskinesias, their functional effects, the quantity of time spent in the off state, the functional effects of motor variations, and their level of complexity.

Finding and validating biomarkers for Parkinson's disease diagnostics, prognosis, and disease progression is a challenge. UDRS scores are given to the patients by the physician. Although this method is subjective, but it has been used in past research. According to research, protein or peptide abnormalities are a major factor in both the development and progression of Parkinson's disease. Therefore, in this work, we predict the UPDRS score, which provides information on the course of the illness, using protein and peptide data collected from individuals with Parkinson's disease. According to the past literature (Chelliah et al. 2022; Ozgul et al. 2015; Ishigami et al. 2012; Yadav et al. 2022), we know that people have worked on proteins and peptides features. Using this data, a lot of significant discoveries have been made, yet there are currently no conclusive biomarkers or treatments.

Motivated by the past study, The major goal of the present research is to use changes in protein and peptide levels over time in Parkinson's disease patients to create an automated diagnostic tool that can forecast UPDRS score.

1.3. Objectives

The current study focuses on the following:

1. Predict “Unified Parkinson's Disease Rating Scale” (UPDRS) scores, which track Parkinson's disease development, using protein and peptide data collected from individuals with the disease.
2. To explore the efficiency of popular machine learning algorithms to build an automated diagnostic tool for predicting updrs scores using Uniprot and Peptides.
3. To identify the most important features(Uniprot and Peptide) for predicting progression of Parkinson’s disease.

2. Literature Review

This section focuses on some recent research utilizing protein and peptide data as biomarkers to identify progression of Parkinson's disease (PD). In literature (Yadav et al. 2022; Chelliah et al. 2022; Ishigami et al. 2012; Ozgul et al. 2015), research work was conducted to diagnose Parkinson's disease using protein and peptide data as biomarkers.

(Yadav et al. 2022) aimed to identify biomarkers for Parkinson's disease (PD) using miRNomics, proteomics, and bioinformatics approaches. Unbiased sets of miRNAs and proteins from blood samples were examined using a rat model that had been exposed to rotenone. For global protein profiling, “liquid chromatography-tandem mass spectrometry” (LC-MS/MS) was utilised, whereas the OpenArray framework was utilised for high volumes miRNA profiling. The bioinformatics research included inscription, functional classification, functional development, analysis of networks, and analysis of the miRNA-protein interaction. In the blood of rotenone-exposed rats, 96 proteins and 19 miRNAs showed considerable upregulation, whereas 22 proteins showed significant downregulation. A network study found many pathways connected to PD. The research discovered possible biomarkers that might be used to diagnose PD, including miR-144, miR-96, miR-29a, PLP1, TUBB4A, and TUBA1C. However, these biomarkers need to be further validated in bigger studies including human populations.

(Ishigami et al. 2012) a proteomic sequencing approach was developed for the differential diagnosis of Parkinson's disease (PD) and "multiple system atrophy" (MSA). As controls, cerebrospinal fluid samples were taken from people who had PD, MSA, and other neurological conditions. Both mass spectrometry analysis and the use of magnetic beads to enrich peptides and proteins from cerebrospinal fluid were used. In order to categorise diseases, data dimension reduction and feature selection were carried out using principal component analysis and support vector machine approaches. The proteome profiles of PD, MSA, and controls were successfully distinguished from one another, and an appropriate classification was established using a support vector machine classifier. The peak at m/z 6250 was shown to be crucial for distinguishing MSA from PD, particularly in the early stages. Especially in the early stages, our proteomic pattern classification technique shows potential for enhancing the precision of clinical diagnosis for PD and MSA.

(Ozgul et al. 2015) concentrate on characterizing two mutant versions of the Parkin protein (Q311R and A371T) associated with Parkinson's disease using biochemical and proteomic techniques. Induced cell lines were created in neuroblastoma cells that expressed Parkin proteins in both their wild-type and mutant forms. 13 of the 22 differently regulated proteins were shown to be uniquely altered in cells expressing mutant Parkin, according to the 2D-DIGE proteomic study. These altered proteins were mostly linked to energy metabolism and protein folding, suggesting that Parkinson's disease may be relevant to them. Both wild-type and mutant Parkin proteins displayed biological activity and comparable stability, but they varied in post-translational modification and susceptibility to protease cleavage, according to biochemical characterisation. This research sheds important light on the potential relationship between certain Parkin mutations, proteome changes, and Parkinson's disease.

(Chelliah et al. 2022) focused on analyzing blood-based candidate biomarkers for Parkinson's disease (PD) identified through proteomic platforms. Following a review of 12 controlled PD trials, 115 potential biomarkers were uncovered, of which 23 were determined to be repeatable across different cohorts. Apolipoprotein A-I (ApoA-I) has been recognised as a highly reproducible biomarker that consistently exhibits downregulation in different cohorts.

ApoA-I contributes to oxidative stress, neuroprotection, and lipid metabolism. Its link to oxidative stress, statins, and cholesterol is explored, along with its potential as a PD biomarker for timely identification and prediction of disease progression. This work uses proteome analysis to suggest ApoA-I as a possible biomarker for PD.

3. Materials and Methods

3.1. Dataset Details

In the present study, a dataset from Kaggle online competition “AMP-Parkinson's Disease Progression prediction” is used. Kaggle is one of the largest and most popular platforms for data science competitions and collaboration. This competition's objective is to forecast MDS-UPDR scores, which reflect the rate of progression of Parkinson's disease in individuals. The public-private collaboration that serves as the competition's host, the Accelerating Medicines collaboration® Parkinson's Disease (AMP®PD), is run by the Foundation of the “National Institutes of Health” (FNIH).(“AMP®-Parkinson’s Disease Progression Prediction | Kaggle”).

There are mainly 3 csv files, which are : tain_peptides, tain_proteins and train_clinical.

tain_peptides `(981834,6)` file consist following attributes:

1. ‘visit_id’ : For the visit, an ID number.
2. ‘visit_month’ : The visit's month.
3. ‘patient_id’ : Patient identification number.
4. ‘UniProt’ : The protein's associated UniProt ID code.
5. ‘Peptide’ : The peptide's amino acid composition in order
6. ‘PeptideAbundance’ : The frequency of the amino acid in the sample.

tain_proteins `(232741,5)` file consist following attributes:

1. ‘visit_id’ : For the visit, an ID number.
2. ‘visit_month’ : The visit's month.
3. ‘patient_id’ : Patient identification number.
4. ‘UniProt’ : The protein's associated UniProt ID code.
5. ‘NPX’ : How often the protein appears in the sample.

tain_clinical `(2615,8)` file consist following attributes:

1. ‘visit_id’ : For the visit, an ID number.
2. ‘visit_month’ : The visit's month.
3. ‘patient_id’ : Patient identification number.
4. ‘updrs_(1-4)’ : The patient's score on the UPDRS's component N.
5. ‘upd23b_clinical_state_on_medication’ : Whether the patient was taking any medication at the time of the UPDRS evaluation, such as Levodopa.

The details of the train_clinical dataset are mentioned in Table 1.

Table 1. Details of Clinical Dataset used in the work

Visits (0-108)	No of subjects (2615)	updrs_1		updrs_2		updrs_3		updrs_4	
		min-max	m ± std	min-max	m ± std	min-max	m ± std	min-max	m ± std
0	248	0-20	5.6±4.6	0-26	4.4±4.9	0-52	13.7±11.9	0-11	2.0±3.2
3	115	0-21	5.5±4.6	0-23	6.6±5.2	1-45	20.5±9.7	0-1	0.1±0.4
6	192	0-33	7.1±5.7	0-28	6.9±5.9	0-56	20.3±13.2	0-17	2.3±3.9
9	99	0-22	6.1±4.8	0-28	7.2±5.5	0-61	20.1±11.1	0-6	0.5±1.4
12	243	0-28	6.2±4.9	0-24	5.3±5.5	0-68	16.2±14.3	0-13	1.3±2.6
18	187	0-28	7.3±5.2	0-25	6.9±5.8	0-62	19.0±13.0	0-13	1.2±2.6
24	243	0-25	6.7±5.2	0-24	5.6±5.7	0-60	16.5±15.2	0-14	1.6±3.0
30	173	0-27	8.2±5.3	0-26	7.6±6.0	0-64	21.7±13.8	0-14	1.7±3.2
36	226	0-27	7.3±5.6	0-26	6.2±6.3	0-67	18.3±16.0	0-20	1.7±3.2
42	154	0-29	8.3±5.7	0-26	8.3±6.6	0-68	22.7±14.1	0-15	1.8±2.9
48	196	0-28	7.6±6.2	0-28	7.1±7.1	0-78	19.3±16.5	0-14	1.9±3.0
54	110	0-31	8.6±6.5	0-39	9.8±7.2	0-86	26.3±13.8	0-14	2.0±3.0
60	166	0-31	7.3±6.1	0-40	6.5±6.9	0-85	19.8±17.6	0-16	2.2±3.1
72	93	0-26	8.4±5.4	0-27	8.8±6.1	0-56	26.4±15.3	0-11	2.3±2.6
84	100	0-28	7.7±5.6	0-36	8.2±7.8	0-66	22.6±18.4	0-10	2.9±2.7
96	58	0-23	7.7±5.5	0-27	7.8±8.1	0-60	21.3±21.0	0-11	4.2±2.9
108	12	1-31	9.4±8.8	0-28	8.2±9.1	0-47	25.2±21.6	0-6	2.6±2.7

*m ± std : mean ± standard deviation.

From the given datasets, we prepared three datasets (protein, peptide, and protein_peptide) on which we worked in our current study. We flattened the datasets train_protiens and train_peptides with respect to the Uniport and Peptide attributes, respectively. And for the protein_pepdite dataset, we merged both protein and peptide flattened data. Then the values

of updrs_scores were merged into all three flattened datasets from train_clinical based on visit_id. Then, for updrs_i, drop all updrs_j except i; after that, we drop the columns and rows containing more than 15% Nan values. And finally, drop the rows for which the value of updrs_i is Nan. After all these steps, we now have all three datasets on which we worked in this project. Details of the used datasets are given in Table 2 below.

Table 2. Details of Final Dataset used in the work

Datasets	updrs_1		updrs_2		updrs_3		updrs_4	
	m	n	m	n	m	n	m	n
Protein	894	177	894	177	884	177	514	173
Peptide	889	769	879	769	879	769	435	743
Protien_Peptide	890	945	890	945	880	945	251	915

*m : number of samples and n : number of features.

3.2. Pre-processing

The pre-processing involves the following operations:

1. We flattened the datasets train_protiens and train_peptides with respect to the Uniport and Peptide attributes, respectively.
   Protein = 1113*228 (1113=unique visit id, 227(unique uniport) + 1(visit_id))
   Peptide = 1113*969 (1113=unique visit id, 968(unique peptides) + 1(visit_id))
2. We merged both protein and peptide flattening data.
   protein_peptide = 1113*1196 (1113=unique visit id, 1196 = 227(unique uniport) +968(unique peptides) + 1(visit_id))
3. Then the values of updrs_scores were merged into all three flattened datasets from train_clinical based on visit_id.
4. For updrs_i, drop all updrs_j except i.
5. Drop the columns and rows containing more than 15% Nan values.
6. Drop the rows for which the value of updrs_i is Nan.
7. Fill the remaining Nan values with 0.
8. Divide the dataset into train and test and after that normalize the train data (except updrs_i) and with the help of this normalize the test data.

After all these points of preprocessing, the final dimension of all four datasets for each updrs_score is given in Table 2.

3.3. Proposed Model

3.3.1. Regression

The next step was to learn a Regression model.In the current study, the training dataset was trained using Support Vector Regression, Random Forest Regression, and MultiLayer Perceptron machine learning models. Since the size of the dataset was large thus, Regression performance was assessed using the K-Fold Cross-Validation technique.

Using the K-fold Cross-validation approach, the input dataset is separated into K groups of equal-sized samples. These specimens are known as folds. For each learning set, the prediction function uses k-1 folds, and for each test set, it uses the remaining folds. This method is a widely used strategy for CVs since it is straightforward and produces findings that are less biased than those of other methods.

A brief description of utilized machine learning methods is as follows:

1) Support Vector Regression

A variation of the popular Support Vector Machines (SVM) method used to solve regression problems is called Support Vector Regression (SVR). SVR concentrates on fitting the data within a defined range known as the "epsilon-insensitive tube," as opposed to classic regression algorithms that try to minimise the error between the predicted and observed values. A versatile method for doing regression problems, SVR is resilient to outliers and may capture nonlinear correlations.

2) Random Forest Regression

Random Forest Regression is a machine learning algorithm that utilizes an ensemble of decision trees to perform regression tasks. It's an expansion of the Random Forest technique, which is mostly used for classification problems. For continuous target variables, Random Forest Regression combines the predictions of several decision trees to provide more precise and reliable predictions. For regression problems, Random Forest Regression is a flexible

and effective technique. Complex interactions may be handled, noise and outliers are robustly handled, and a measure of feature relevance is provided.

3) MultiLayer Perceptron

Multilayer Perceptrons (MLPs) are a form of artificial neural network (ANN) often used for different machine learning applications, including classification, regression, and pattern recognition. Regression tasks using Multilayer Perceptrons (MLPs) aim to predict a continuous objective variable. MLP regression entails changing an MLP's architecture and loss function to accommodate regression issues.A versatile and effective method for modeling and projecting continuous target variables is provided by MLP regression. MLP regression models may successfully capture complicated connections and generate precise predictions in regression tasks when the appropriate architectural design, activation functions, loss functions, and hyperparameter tuning are used.

Overall, in this work, we have used UniProt and Peptide values as features to predict the values of updrs scores. For this we have used machine learning algorithms to train a Regression model. The pipeline of the trained model is shown in Figure 1.

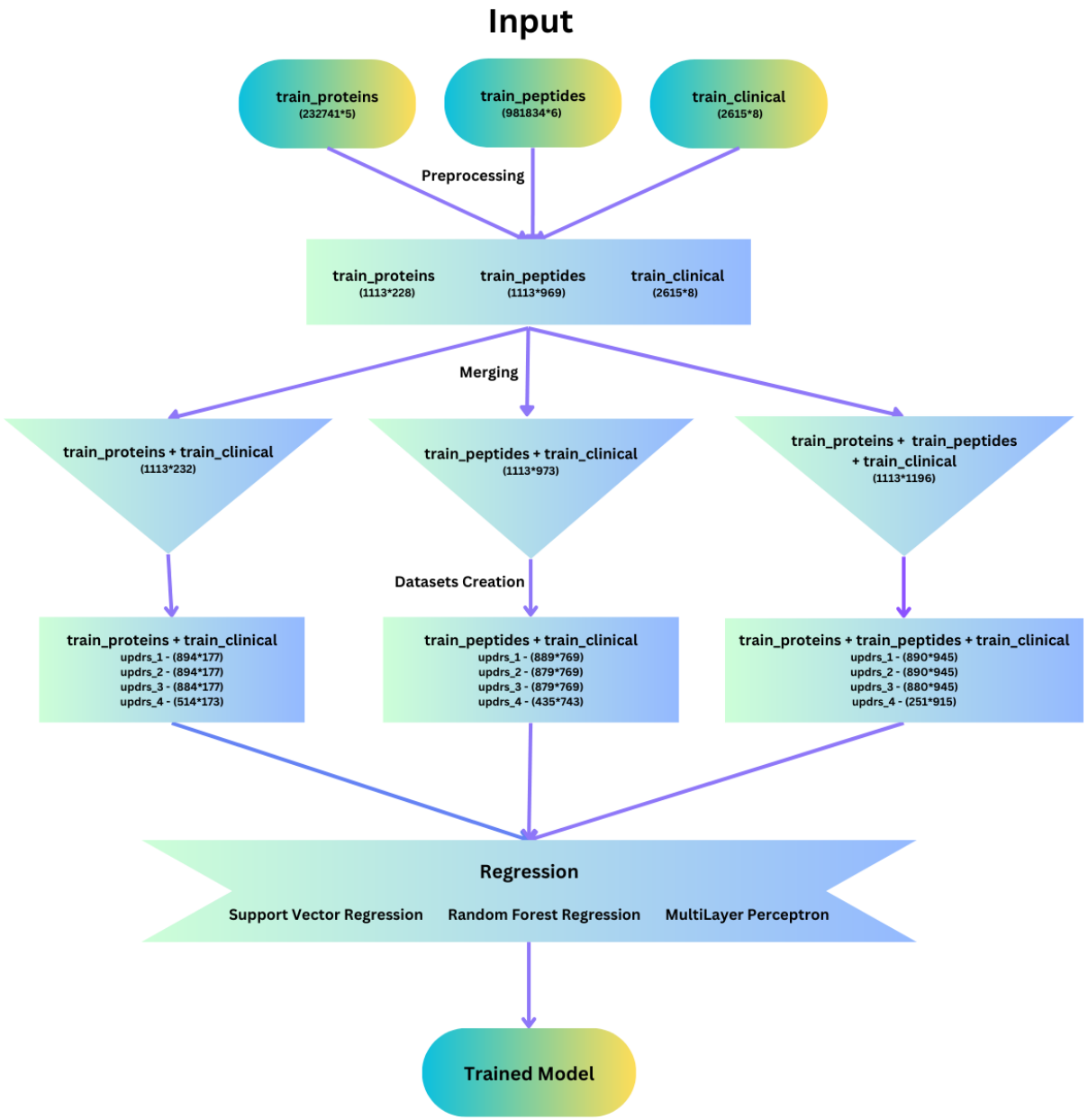

Figure 1. The pipeline for training a regression model

4. Experimental Setup & Results

The whole present study was conducted using the COLAB platform. Libraries used in COLAB are Pandas, Numpy, Matplotlib, and Sklearn.("Google Colaboratory").

4.1. Regression results

The experiments were performed using the K-Fold Cross Validation approach and the performance is measured using Mean Squared Error(MSE), Root Mean Squared Error(RMSE), and R_Score(r2) measures. The average performance of three regressors is reported in Tables 3, 5, and 7 and Table 4, 6, and 8 shows the rank of the models according to their performance. For each regressor performance is reported as the mean value of 5 folds ± standard deviation in 5 folds.

Table 3: Regression results for protein features

Models	Performanse	updrs_1	updrs_2	updrs_3	updrs_4
SVR	MSE	25.34±0.07	31.53±0.09	207.42±1.00	9.68±0.05
	RMSE	4.99±0.01	5.59±0.01	14.35±0.04	3.05±0.02
	R2	0.06±0.00	0.07±0.00	0.07±0.01	-0.11±0.01
RF	MSE	23.40±0.08	26.42±0.33	166.20±1.69	8.03±0.18
	RMSE	4.81±0.01	5.11±0.02	12.86±0.06	2.79±0.03
	R2	0.12±0.01	0.22±0.00	0.25±0.01	0.04±0.03
MLP	MSE	25.1±0.45	26.42±0.33	172.16±4.45	8.64±0.08
	RMSE	4.98±0.04	5.11±0.02	13.08±0.17	2.91±0.01
	R2	0.06±0.02	0.22±0.00	0.22±0.02	-0.04±0.04

Table 4: Performance Rank of models for protein features

	SVR	RF	MLP
updrs_1	2	1	2
updrs_2	3	1	1
updrs_3	3	1	2
updrs_4	3	1	2

Table 5: Regression results for peptide features

Models	Performanse	updrs_1	updrs_2	updrs_3	updrs_4
	MSE	25.58±0.08	30.85±0.11	201.61±0.40	9.14±0.04
	RMSE	5.02±0.00	5.51±0.01	14.16±0.022	2.96±0.02
SVR	**R2**	0.04±0.00	0.07±0.00	0.07±0.00	-0.15±0.00
	MSE	22.56±0.04	26.29±0.28	162.98±1.14	7.95±0.19
	RMSE	4.71±0.01	5.10±0.02	12.73±0.03	2.82±0.02
RF	**R2**	0.51±0.00	0.20±0.01	0.25±0.01	0.085±0.02
	MSE	20.54±1.15	22.85±0.67	132.97±4.15	7.54±0.18
	RMSE	4.49±0.10	4.74±0.05	11.47±0.17	2.71±0.04
MLP	**R2**	0.22±0.04	0.30±0.01	0.38±0.02	0.01±0.05

Table 6: Performance Rank of models for peptide features

	SVR	RF	MLP
updrs_1	3	1	2
updrs_2	3	2	1
updrs_3	3	2	1
updrs_4	3	1	2

Table 7: Regression results for protein and peptide features

Models	Performanse	updrs_1	updrs_2	updrs_3	updrs_4
	MSE	25.29±0.08	30.86±0.10	198.49±0.20	7.52±0.03
	RMSE	5.00±0.02	5.51±0.03	14.04±0.02	2.66±0.03
SVR	**R2**	0.05±0.00	0.07±0.00	0.08±0.00	-0.13±0.01
	MSE	22.80±0.18	25.73±0.13	160.24±1.43	6.63±0.15
	RMSE	4.74±0.02	5.05±0.01	12.61±0.05	2.52±0.03
RF	**R2**	0.14±0.01	0.22±0.00	0.25±0.01	-0.04±0.03
	MSE	28.13±0.97	34.90±2.17	126.40±3.89	6.43±0.23
	RMSE	5.27±0.07	5.87±0.17	11.17±0.16	2.49±0.05
MLP	**R2**	-0.06±0.03	-0.06±0.07	0.41±0.02	-0.03±0.05

Table 8: Performance Rank of models for protein and peptide features

	SVR	RF	MLP
updrs_1	2	1	3
updrs_2	2	1	3
updrs_3	3	2	1
updrs_4	3	2	1

The following can be observed from the Tables 3 to 8:

- The maximum r_score of 0.51±0.00 with Random Forest for updrs_1 by peptide features is achieved.
- For updrs_2 the maximum r_score of 0.30±0.01 with MLP by peptide features is achieved. The r_score of 0.22±0.00 with Random Forest by protein and protien_peptide features is achieved. However, peptide features gave the result with a smaller number of features in comparison to protien_peptide features.
- For updrs_3 the maximum r_score of 0.41±0.02 with MLP by protein_peptide features is achieved. Using peptide features, r_score of 0.38±0.02 with MLP is achieved.
- For updrs_4 the maximum r_score of 0.085±0.02 with Random Forest by peptide features is achieved.
- We get good results either by RF or by MLP, these two regressors doing well.

4.2. Identification of important features

For each updrs score, we backtracked the features which contribute maximum in predicting the values of updrs score. We computed the impotent features, and they are reported here.

For updrs_1

The most important features for updrs_1 are 'Q06481', 'P61916', 'P17174', 'GEAGAPGEEDIQGPTK', 'P14618', 'LEEQAQQIR', 'NVVYTC(UniMod_4)NEGYSLIGNPVAR', 'QQTHMLDVMQDHFSR', 'P05067', 'P07602', 'TQSSLVPALTDFVR', 'P04180', 'O60888', 'Q6UXB8', 'FFLC(UniMod_4)QVAGDAK', 'NLREGTC(UniMod_4)PEAPTDEC(UniMod_4)KPVK', 'P05060', 'RYIETDPANRDR',

'WYFDVTEGK', 'AVLPTGDVIGDSAK', 'GATLALTQVTPQDER', 'LPYTASSGLMAPR', 'NTGIIC(UniMod_4)TIGPASR', 'P08571', 'IGADFLAR', 'KC(UniMod_4)STSSLLEAC(UniMod_4)TFR', 'LGC(UniMod_4)SLNQNSVPDIHGVEAPAR', 'LMVELHNLYR', 'LRENELTYYC(UniMod_4)C(UniMod_4)KK', 'QQETAAAETETR', 'RPGGEPSPEGTTGQSYNQYSQR', 'AATVGSLAGQPLQER', 'DTVIKPLLVEPEGLEK', 'EVNVSPC(UniMod_4)PTQPC(UniMod_4)QLSK', 'GYPGVQAPEDLEWER', 'ITTTSPWMFPSR', 'LGPLVEQGRVR', 'LVWEEAMSR', 'O00533', 'P02751', 'P13611', 'Q15904', 'QYNVGPSVSKYPLR', 'YQC(UniMod_4)YC(UniMod_4)YGR', 'AANEVSSADVK', 'AIGYLNTGYQR', 'ALEYIENLR', 'ASNLESGVPSR', 'ATEDEGSEQKIPEATNR', 'AYQGVAAPFPK', 'DTINLLDQR', 'EILSVDC(UniMod_4)STNNPSQAK', 'FEHC(UniMod_4)NFNDVTTR', 'FSPATHPSEGLEENYC(UniMod_4)RNPDNDPQGPWC(UniMod_4)YTTDPEKR', 'FTQVTPTSLSAQWTPPNVQLTGYR', 'GC(UniMod_4)SFLPDPYQK', 'IALVITDGR', 'LKDDEVAQLKK', 'MDASLGNLFAR', 'NTFAEVTGLSPGVTYYFK', 'O14773', 'P01857', 'P02748', 'P08123', 'P09486', 'P09871', 'P13521', 'P16870', 'P23142', 'P43121', 'P98160', 'Q8NBJ4', 'Q92823', 'QHVVYGPWNLPQSSYSHLTR', 'QWAGLVEK', 'SC(UniMod_4)DKTHTC(UniMod_4)PPC(UniMod_4)PAPELLGGPSVFLFPPKPK', 'SLPSEASEQYLTK', 'SNSSMHITDC(UniMod_4)R', 'THPHFVIPYR', 'TM(UniMod_35)QENSTPRED', 'TYLGNALVC(UniMod_4)TC(UniMod_4)YGGSR', 'VLSIAQAHSPAFSC(UniMod_4)EQVR', 'VRQGQGQSEPGEYEQR', 'WELALGR', 'WSGQTAIC(UniMod_4)DNGAGYC(UniMod_4)SNPGIPIGTR', and 'YHDRDVWKPEPC(UniMod_4)R'.

For updrs_2

The most important features for updrs_2 are 'MKYWGVASFLQK', 'Q06481', 'LQDLYSIVR', 'P04180', 'P02753', 'SSGLVSNAPGVQIR', 'LEEQAQQIR', 'DQGNQEQDPNISNGEEEEEKEPGEVGTHNDNQER', 'ITTTSPWMFPSR', 'NLNEKDYELLC(UniMod_4)LDGTR', 'YWGVASFLQK', 'ISYGNDALMPSLTETK', 'AKLEEQAQQIR', 'GEVQAMLGQSTEELR', 'KC(UniMod_4)STSSLLEAC(UniMod_4)TFR', 'P13521', 'FFLC(UniMod_4)QVAGDAK', 'P05060', 'FEHC(UniMod_4)NFNDVTTR', 'GYPGVQAPEDLEWER', 'LIVHNGYC(UniMod_4)DGR', 'NLQPASEYTVSLVAIKGNQESPK', 'SPFEQHIK',

'AATVGSLAGQPLQER', 'ALEYIENLR', 'C(UniMod_4)AEENC(UniMod_4)FIQK', 'LEAGDHPVELLAR', 'LMVELHNLYR', 'LQAEAFQAR', 'O15240', 'P02649', 'P05067', 'P13987', 'P14618', 'Q14118', 'QQTHMLDVMQDHFSR', 'TPLGDTTHTC(UniMod_4)PR', 'VFQEPLFYEAPR', 'WELALGR', 'AKPALEDLR', 'ALFLETEQLK', 'ALPGTPVASSQPR', 'ASNLESGVPSR', 'AYKSELEEQLTPVAEETR', 'DQTVSDNELQEMSNQGSK', 'DRLDEVKEQVAEVR', 'ELLESYIDGR', 'GLEFLSVPSTYYK', 'GQEDASPDRFSNTDYAVGYMLR', 'IEEELGDEAR', 'IEIPSSVQQVPTIIK', 'IGADFLAR', 'IMNGEADAMSLDGGFVYIAGK', 'IYLYTLNDNAR', 'KAEEEHLGILGPQLHADVGDKVK', 'KLGQSLDC(UniMod_4)NAEVYVVPWEK', 'KPVDEYKDC(UniMod_4)HLAQVPSHTVVAR', 'KTLLSNLEEAKK', 'LAAAVSNFGYDLYR', 'LDELRDEGK', 'LDIDSPPITAR', 'LGQSLDC(UniMod_4)NAEVYVVPWEK', 'LPPTSAHGNVAEGETKPDPDVTER', 'LPYTASSGLMAPR', 'MATLYSR', 'MC(UniMod_4)PQLQQYEMHGPEGLR', 'MYLGYEYVTAIR', 'NFGYTLR', 'NLREGTC(UniMod_4)PEAPTDEC(UniMod_4)KPVK', 'NTFAEVTGLSPGVTYYFK', 'O14791', 'P01608', 'P02751', 'P02787', 'P02790', 'P04156', 'P04207', 'P07602', 'P08123', 'P09104', 'P39060', 'P43121', 'P61916', 'Q6UXB8', 'Q92823', 'QRQEELC(UniMod_4)LAR', 'QWAGLVEK', 'SASDLTWDNLKGK', 'SRYPVC(UniMod_4)GSDGTTYPSGC(UniMod_4)QLR', 'SSNLIILEEHLK', 'SSQGGSLPSEEK', 'TALASGGVLDASGDYR', 'TGAQELLR', 'VLLDGVQNPR', 'VQAAVGTSAAPVPSDNH', 'VVEQMC(UniMod_4)ITQYER', 'WC(UniMod_4)AVSEHEATK', and 'YQC(UniMod_4)YC(UniMod_4)YGR'.

For updrs_3

The most important features for updrs_3 are 'DQGNQEQDPNISNGEEEEEKEPGEVGTHNDNQER', 'O15240', 'P10645', 'YPGPQAEGDSEGLSQGLVDREK', 'P13521', 'ALEYIENLR', 'P02753', 'KPQSAVYSTGSNGILLC(UniMod_4)EAEGEPQPTIK', 'O00533', 'P01780', 'P05067', 'C(UniMod_4)LVEKGDVAFVKHQTVPQNTGGK', 'GYPGVQAPEDLEWER', 'IEIPSSVQQVPTIIK', 'Q14515', 'TLKIENVSYQDKGNYR', 'MKYWGVASFLQK', 'VNGSPVDNHPFAGDVVFPR', 'YWGVASFLQK', 'AYQGVAAPFPK', 'DVQLVESGGGLVKPGGSLR', 'EFQLFSSPHGK', 'GEAGAPGEEDIQGPTK', 'IASFSQNC(UniMod_4)DIYPGKDFVQPPTK', 'KLSENTDFLAPGVSSFTDSNQQESITK', 'SEALAVDGAGKPGAEEAQDPEGK', 'DC(UniMod_4)HLAQVPSHTVVAR',

'FTILDSQGK', 'KC(UniMod_4)STSSLLEAC(UniMod_4)TFR', 'LQDLYSIVR', 'LRTEGDGVYTLNNEK', 'P01042', 'P05452', 'Q06481', 'Q92520', 'QHVVYGPWNLPQSSYSHLTR', 'ALDFAVGEYNK', 'ALGISPFHEHAEVVFTANDSGPRR', 'ARAEAQEAEDQQAR', 'ASYLDC(UniMod_4)IR', 'AYKSELEEQLTPVAEETR', 'DALSSVQESQVAQQAR', 'FDEFFSEGC(UniMod_4)APGSKK', 'GNPEPTFSWTK', 'GVASLFAGR', 'ILAGSADSEGVAAPR', 'IQPSGGTNINEALLR', 'ISLPESLK', 'LEGQEEEEDNRDSSMK', 'LEPGQQEEYYR', 'LIADLGSTSITNLGFR', 'LIVHNGYC(UniMod_4)DGR', 'LLPAQLPAEKEVGPPLPQEAVPLQK', 'LNMHMNVQNGKWDSDPSGTK', 'LRTEGDGVYTLNNEKQWINK', 'LSINTHPSQKPLSITVR', 'LVFFAEDVGSNK', 'MASGAANVVGPK', 'NSLFEYQK', 'NTFAEVTGLSPGVTYYFK', 'NVVYTC(UniMod_4)NEGYSLIGNPVAR', 'O15394', 'P00738', 'P02675', 'P02787', 'P04156', 'P05060', 'P43121', 'P55290', 'Q08380', 'Q14118', 'Q92823', 'Q96KN2', 'Q99829', 'Q99832', 'Q9NYU2', 'QQTHMLDVMQDHFSR', 'QRQEELC(UniMod_4)LAR', 'QVVAGLNFR', 'QWAGLVEK', 'SIVVSPILIPENQR', 'SLNNQIETLLTPEGSR', 'SSQGGSLPSEEK', 'STNLHDYGMLLPC(UniMod_4)GIDK', 'TEGDGVYTLNNEK', 'VQAAVGTSAAPVPSDNH', 'VRQGQGQSEPGEYEQR', 'WC(UniMod_4)AVSEHEATK', 'WYFDVTEGK', 'YGLVTYATYPK', 'YPSLSIHGIEGAFDEPGTK', and 'YQC(UniMod_4)YC(UniMod_4)YGR'.

For updrs_4

The most important features for updrs_4 are 'HLSLLTTLSNR', 'YHDRDVWKPEPC(UniMod_4)R', 'GYC(UniMod_4)APGMEC(UniMod_4)VK', 'STNLHDYGMLLPC(UniMod_4)GIDK', 'RGEQC(UniMod_4)VDIDEC(UniMod_4)TIPPYC(UniMod_4)HQR', 'KVPQVSTPTLVEVSR', 'DSGEGDFLAEGGGVR', 'P00736', 'QKVEPLRAELQEGAR', 'QQTHMLDVMQDHFSR', 'SC(UniMod_4)SPELQQK', 'AAFTEC(UniMod_4)C(UniMod_4)QAADK', 'P01857', 'Q15904', 'C(UniMod_4)C(UniMod_4)AAADPHEC(UniMod_4)YAK', 'EAEEETTNDNGVLVLEPARK', 'VTGVVLFR', 'P07225', 'P23083', 'INHC(UniMod_4)RFDEFFSEGC(UniMod_4)APGSKK', 'KAADDTWEPFASGK', 'LC(UniMod_4)MGSGLNLC(UniMod_4)EPNNKEGYYGYTGAFR', 'LKC(UniMod_4)DEWSVNSVGK', 'NVVYTC(UniMod_4)NEGYSLIGNPVAR', 'O00584',

'P02656', 'P02790', 'P16152', 'P24592', 'AADDTWEPFASGK', 'AAFGQGSGPIMLDEVQC(UniMod_4)TGTEASLADC(UniMod_4)K', 'C(UniMod_4)LAPLEGAR', 'EGQEC(UniMod_4)GVYTPNC(UniMod_4)APGLQC(UniMod_4)HPPKDDEAPLR', 'EHVAHLLFLR', 'FSPATHPSEGLEENYC(UniMod_4)RNPDNDPQGPWC(UniMod_4)YTTDPEKR', 'GSPAINVAVHVFR', 'IALVITDGR', 'IFSFDGKDVLR', 'KC(UniMod_4)C(UniMod_4)VEC(UniMod_4)PPC(UniMod_4)PAPPVAGPSVFLFPPKPK', 'KFPSGTFEQVSQLVK', 'LPYTASSGLMAPR', 'M(UniMod_35)YLGYEYVTAIR', 'NTGIIC(UniMod_4)TIGPASR', 'P00441', 'P02749', 'P04217', 'P09486', 'P20774', 'P55290', 'P61626', 'Q12841', 'QAPGQGLEWMGR', 'SASDLTWDNLK', 'SDVMYTDWKK', 'TSPVDEKALQDQLVLVAAK', 'VDGALC(UniMod_4)MEK', 'VFSNGADLSGVTEEAPLKLSK', 'ALMSPAGMLR', 'ALSSEWKPEIR', 'AVLPTGDVIGDSAK', 'C(UniMod_4)C(UniMod_4)TESLVNR', 'C(UniMod_4)KPVNTFVHEPLVDVQNVC(UniMod_4)FQEK', 'C(UniMod_4)PFPSRPDNGFVNYPAKPTLYYK', 'DSGFQMNQLR', 'DSGVPDRFSGSGSGTDFTLK', 'DVQLVESGGGLVKPGGSLR', 'ELSSFIDKGQELC(UniMod_4)ADYSENTFTEYK', 'EQLSLLDRFTEDAKR', 'FKDLGEENFK', 'GC(UniMod_4)SFLPDPYQK', 'GIYGTISR', 'GLYDVVSVLR', 'GNSYFMVEVK', 'HGTC(UniMod_4)AAQVDALNSQKK', 'HYDGSYSTFGER', 'IEIPSSVQQVPTIIK', 'IEKVEHSDLSFSK', 'IKPVFIEDANFGR', 'ILEVVNQIQDEER', 'ILTC(UniMod_4)M(UniMod_35)QGMEEIR', 'ISLPESLK', 'KAEEEHLGILGPQLHADVGDKVK', 'KDSGFQM(UniMod_35)NQLR', 'KSQPMGLWR', 'KVESELIKPINPR', 'KVLLDGVQNPR', 'LC(UniMod_4)TVATLR', 'LDELRDEGK', 'LEPGQQEEYYR', 'LETPDFQLFK', 'LGMDGYR', 'LKC(UniMod_4)DEWSVNSVGKIEC(UniMod_4)VSAETTEDC(UniMod_4)IAK', 'LQDLYSIVR', 'LSELIQPLPLER', 'LSINTHPSQKPLSITVR', 'LVNEVTEFAK', 'MDASLGNLFAR', 'NEQEQPLGQWHLS', 'NILTSNNIDVK', 'NQEQVSPLTLLK', 'NRDHDTFLAVR', 'O14773', 'O15394', 'P02766', 'P02774', 'P04156', 'P04211', 'P05546', 'P39060', 'P43652', 'Q96PD5', 'Q9NQ79', 'Q9UBX5', 'QC(UniMod_4)VPTEPC(UniMod_4)EDAEDDC(UniMod_4)GNDFQC(UniMod_4)STGR', 'RGYQLSDVDGVTC(UniMod_4)EDIDEC(UniMod_4)ALPTGGHIC(UniMod_4)SYR', 'RLEAGDHPVELLAR', 'RPDSLQHVLLPVLDR', 'SASDLTWDNLKGK', 'SIVVSPILIPENQR', 'SPAINVAVHVFR',

'SRYPVC(UniMod_4)GSDGTTYPSGC(UniMod_4)QLR', 'SSALDMENFR', 'SVPMVPPGIK', 'TELLPGDRDNLAIQTR', 'TQVNTQAEQLR', 'TVAAC(UniMod_4)NLPIVR', 'VHVSEEGTEPEAMLQVLGPKPALPAGTEDTAKEDAANRK', 'VVEQMC(UniMod_4)ITQYER', 'YIVSGTPTFVPYLIK', 'YQC(UniMod_4)YC(UniMod_4)YGR', 'YVGGQEHFAHLLILR', and 'YYTYLIMNK'.

5. Conclusion & Future Work

In this work, we developed an automated system to predict updrs_score based on machine learning. We constructed four datasets containing different numbers of features from the datasets given on Kaggle online competition "AMP-Parkinson's Disease Progression prediction". We investigated three regressors, namely Support Vector Regression, Random Forest Regression, and MultiLayer Perceptron. Features (Uniport and peptides) are used for training a regression model. Also, a subset of relevant features is recorded.

The best r_score of 0.51±0.00 was achieved with peptide features using Random Forest for updrs_1. The important features are the QQTHMLDVMQDHFSR, Q06481, NVVYTC(UniMod_4)NEGYSLIGNPVAR, P05067, P05060, LPYTASSGLMAPR, KC(UniMod_4)STSSLLEAC(UniMod_4)TFR, GYPGVQAPEDLEWER, ALEYIENLR, NTFAEVTGLSPGVTYYFK, P13521, P43121, Q92823, QWAGLVEK, Q06481, LQDLYSIVR, KC(UniMod_4)STSSLLEAC(UniMod_4)TFR, P13521, P05060, GYPGVQAPEDLEWER, ALEYIENLR, P05067, IEIPSSVQQVPTIIK, LPYTASSGLMAPR, NTFAEVTGLSPGVTYYFK, P04156, P43121, Q92823, QWAGLVEK, P13521, ALEYIENLR, P05067, GYPGVQAPEDLEWER, IEIPSSVQQVPTIIK, KC(UniMod_4)STSSLLEAC(UniMod_4)TFR, LQDLYSIVR, Q06481, NTFAEVTGLSPGVTYYFK, NVVYTC(UniMod_4)NEGYSLIGNPVAR, P04156, P05060, P43121, Q92823, QWAGLVEK, NVVYTC(UniMod_4)NEGYSLIGNPVAR, LPYTASSGLMAPR, IEIPSSVQQVPTIIK, LQDLYSIVR, and P04156.

References

"AMP®-Parkinson's Disease Progression Prediction | Kaggle." n.d. Accessed July 8, 2023. https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction.

Chelliah, Shalini Sundramurthi, Saatheeyavaane Bhuvanendran, Kasthuri Bai Magalingam, Muhamad Noor Alfarizal Kamarudin, and Ammu Kutty Radhakrishnan. 2022. "Identification of Blood-Based Biomarkers for Diagnosis and Prognosis of Parkinson's Disease: A Systematic Review of Proteomics Studies." Ageing Research Reviews 73 (January): 101514. https://doi.org/10.1016/j.arr.2021.101514.

Goetz, Christopher G., Barbara C. Tilley, Stephanie R. Shaftman, Glenn T. Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, et al. 2008. "Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results: MDS-UPDRS: Clinimetric Assessment." Movement Disorders 23 (15): 2129–70. https://doi.org/10.1002/mds.22340.

"Google Colaboratory." n.d. Accessed July 8, 2023. https://colab.research.google.com/?utm_source=scs-index.

Ishigami, Noriko, Takahiko Tokuda, Masaya Ikegawa, Mika Komori, Takashi Kasai, Takayuki Kondo, Yumiko Matsuyama, et al. 2012. "Cerebrospinal Fluid Proteomic Patterns Discriminate Parkinson's Disease and Multiple System Atrophy." Movement Disorders: Official Journal of the Movement Disorder Society 27 (7): 851–57. https://doi.org/10.1002/mds.24994.

Ozgul, Sinem, Murat Kasap, Gurler Akpinar, Aylin Kanli, Nil Güzel, Kübra Karaosmanoglu, Ahmet Tarik Baykal, and Pervin Iseri. 2015. "Linking a Compound-Heterozygous Parkin Mutant (Q311R and A371T) to Parkinson's Disease by Using Proteomic and Molecular Approaches." Neurochemistry International 85–86: 1–13. https://doi.org/10.1016/j.neuint.2015.03.007.

"Parkinsons Disease and the Ageing Indian Population - Healthcare Radius." 2021. April 14, 2021. https://www.healthcareradius.in/clinical/28890-parkinsons-disease-and-the-ageing-indian-population.

"Parkinson's Disease: Causes, Symptoms, and Treatments." n.d. National Institute on Aging. Accessed January 28, 2023. https://www.nia.nih.gov/health/parkinsons-disease.

Yadav, Sanjeev Kumar, Anuj Pandey, Sana Sarkar, Smriti Singh Yadav, Devendra Parmar, and Sanjay Yadav. 2022. "Identification of Altered Blood MicroRNAs and Plasma Proteins in a Rat Model of Parkinson's Disease." Molecular Neurobiology 59 (3): 1781–98. https://doi.org/10.1007/s12035-021-02636-y.